A Linked Data Profiling Service for Quality Assessment

Nandana Mihindukulasooriya[1], Raúl García-Castro[1], Freddy Priyatna[1], Edna Ruckhaus[1], and Nelson Saturno[2] *

[1] Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain
[2] Universidad Simón Bolívar, Caracas, Venezuela
{nmihindu,rgarcia,fpriyatna,eruckhaus}@fi.upm.es

Abstract. The Linked (Open) Data cloud has been growing at a rapid rate in recent years. However, the large variance of quality in datasets is a key obstacle that hinders their use, so quality assessment has become an important aspect. Data profiling is one of the widely used techniques for data quality assessment in domains such as relational data; nevertheless, it is not so widely used in Linked Data. We argue that one reason for this is the lack of Linked Data profiling tools that are configurable in a declarative manner, and that produce comprehensive profiling information with the level of detail required by quality assessment techniques. To this end, this demo paper presents, Loupe API, a RESTful web service that profiles Linked Data based on user requirements and produces comprehensive profiling information on explicit RDF general data, class, property and vocabulary usage, and implicit data patterns such as cardinalities, instance ratios, value distribution, and multilingualism. Profiling results can be used to assess quality either by manual inspection, or automatically using data validation languages such as SHACL, ShEX, or SPIN.

Keywords: Linked Data, Quality, Data Profiling, Services

1 Introduction

The Linked (Open) Data cloud has been growing at a rapid rate in recent years. Some portions of it come from crowd-sourced knowledge bases such as Wikipedia, while others come from government administrations, research publishers, and other organizations. These datasets have different levels of quality [1] such that for most practical use cases, they need to be assessed to get an indication of their quality.

Juran and Godfrey describe quality using multiple views [2]. On the one hand, quality can be seen as "fit for intended use in operations, decision-making, and planning", i.e., relevance, recency, completeness, and precision. On the other hand, quality is also viewed as "freedom from deficiencies", i.e., correctness and

* This research is partially supported by the 4V (TIN2013-46238-C4-2-R) and MobileAge (H2020/693319) projects and the FPI grant (BES-2014-068449.)

consistency. In either case, quality assessment is needed before using the data for a given task to ensure that the data has an adequate quality level. Further, the results of the assessment can be used to assist the process of improving quality by cleaning and repairing deficiencies in the data. The objective of the work presented in this paper is to provide data profiling service with fine-grained information that can be used as input for many quality assessment tasks related to both these views of data quality.

Detailed data analysis is one common preliminary task in quality assessment, and data profiling is one of the most widely-used techniques for such analysis [3]. Data profiling is defined as the process of examining data to collect statistics and provide relevant metadata about the data [4]. Even though data profiling is widely used in quality assessment in domains such as relational data, we see a lack of usage of data profiling in Linked Data.

In this paper, we describe a Linked Data profiling service, the Loupe API, which provides access to the Loupe tool via a RESTful interface. The Loupe API may be configured to specify the source data as well as the profiling activities it should perform. As a consequence it can be used for different purposes. In the recent years, Loupe has been used to assess the quality of datasets in several projects such as DBpedia [5] and 3Cixty [6]. A RESTful interface facilitates the integration of the Loupe profiling services to other systems. The Loupe API has been integrated with one of our ongoing projects, MappingPedia[1], a collaborative environment for R2RML mappings, in order to gather statistics and do quality assessment since R2RML mappings are themselves RDF/Linked Data datasets.

2 Loupe API

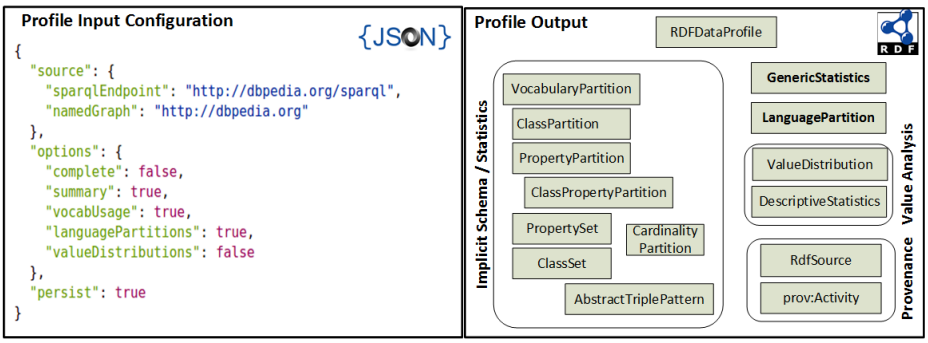
The Loupe API[2] is a configurable Linked Data profiling service. The three main phases in Linked Data profiling are (1) specification of input, (2) execution of data profiling, and (3) representation of profiling results. Listing 1a shows an example input of a Loupe API profile request. Users can specify their requirements (i.e., which profiling tasks to execute) and other configuration details such as how to access the data source (and which data to profile), or whether to persist the profiling results in the Loupe public repository (i.e., they will be available via search). The profiling tasks are grouped into four categories:

- **summary** - provides generic statistics on an RDF data source related to its size and the type of content it has, for example, typed entity count or distinct IRI object count.
- **vocabUsage** - provides information on the implicit schema of the data by analyzing how vocabulary terms such as classes and properties are used, their domains and ranges, cardinalities, uniqueness, among others.
- **languagePartitions** - provides information on multilingual content by analyzing the frequency of each language in language tagged strings.

[1] `http://demo.mappingpedia.linkeddata.es/`
[2] `http://api.loupe.linkeddata.es/`

– **valueDistributions** - provides information on the value distribution of a given property.

The results of profiling are represented in RDF using the Loupe ontology[3]. The main elements of the profiling results are illustrated in Figure 1b; the complete results in RDF are available[4]; we also provide a set of cURL examples[5] for invoking the service.



(a) Input Configuration (b) Output Elements

Fig. 1: A summary of Loupe API input and output

These profiling results can be used for validating the quality of a dataset either by manual inspection or by specifying the validation rules in a language such as SHACL[6], ShEx[7], or SPIN[8]. Data profiling facilitates the manual inspection by providing a high-level summary so that an evaluator can adapt techniques such as exploratory testing [6] to identify strange occurrences in data.

Nevertheless, automatic validation is needed when a large amount of data is present and it is feasible in most situations. For example, data model constraints such as uniqueness of values, expected cardinalities, domains and ranges, inconsistent use of duplicate properties [5] can be easily validated by expressing those in a constraint language and automatically using profiling information.

Further, profiling results enable the analysis of a dataset over a period of time by periodically profiling data and performing the analysis on multiple profiling results. For instance, expected deletions of data or undesired changes can be detected by analysing the changes in the dataset profiles.

The Loupe API is implemented as a RESTful service and currently three operations are available as illustrated in Figure 2.

[3] http://ont-loupe.linkeddata.es/def/core#
[4] https://git.io/vy1t0
[5] https://github.com/nandana/loupe-api/wiki/examples
[6] https://www.w3.org/TR/shacl/
[7] https://shexspec.github.io/spec/
[8] http://spinrdf.org/

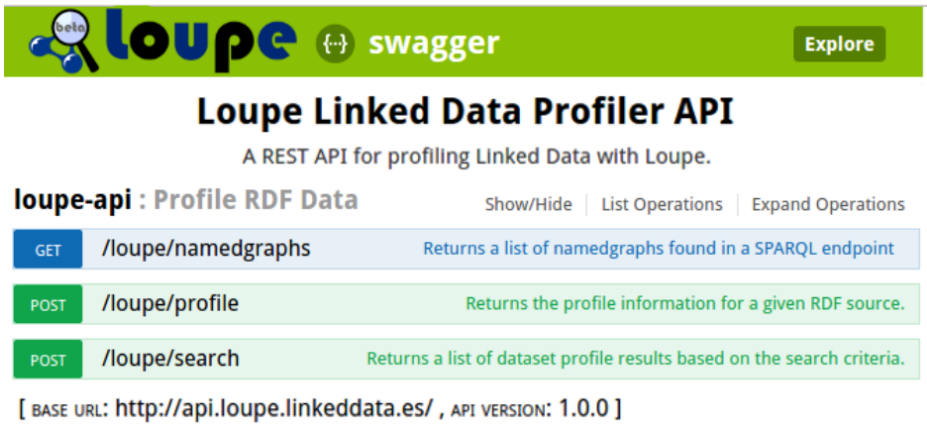


Fig. 2: Loupe API Documentation

3 Related Work

Zaveri et al. [1] present a comprehensive review of data quality assessment techniques and tools in the literature, and propose a conceptual framework with quality metrics grouped in four dimensions: accesibility, intrinsic, contextual, and representational; in particular, it mentions the use of profiling by the ProLOD tool for Semantic Accuracy. The ProLOD tool [7] has a pre-processing clustering and labeling phase and a real-time profiling phase that gathers statistics on a specific cluster in order to detect misused properties and discordant values.

Tools that provide statistics on the Linked Open Data Cloud include Aether [8] that provides extended VOID statistical descriptions of RDF content and interlinking, and ExpLOD [9] and ABSTAT [10] that provide summaries of RDF usage and interlinking.

Differently from the other tools mentioned, the Loupe API is available as a RESTful web service where users can configure and generate Linked Data profiles in RDF using the Loupe ontology. Further, Loupe provides summarized information not only on explicit vocabulary, class and property usage as the other tools but it also facilitates the analysis of implicit data patterns by providing a finer grained set of metrics compared to existing tools, such as instance ratio (ratio of instances of a given class to all entities) and property cardinalities. Low granularity metrics and other capabilities of Loupe have been applied to the analysis of redundant information, consistency with respect to the axioms in the ontology, syntactic validity and detection of outliers.

4 Conclusion and future work

This paper presents the Loupe API, a configurable RESTful service for profiling Linked Data, where results can be used for quality assessment purposes. The

paper illustrated its use, and motivated it with a discussion on how it can be integrated to the quality assessment process.

Nevertheless, there are several challenges in profiling large datasets using a service compared to a standalone tool. Thus, Loupe API is mostly suitable for profiling small datasets. Large datasets (e.g., DBPedia) could take a long time to profile and the requests might timeout. In the future, we plan to provide support for asynchronous executions for such cases.

Another challenge is to detect the capabilities and limitations of the SPARQL endpoint and to adapt to those capabilities. Loupe API uses SPARQL 1.1 features and some metrics are omitted if an endpoint only supports SPARQL 1.0.

In the future, we also plan to extend the profiling service to other Linked Data sources such as RDF dumps, SPARQL construct queries, and LDF endpoints. Further, we plan to allow users to specify their quality requirements in a declarative manner using formal languages such as SHACL, ShEX, SPIN or using an editor with common validation rules. This will allow Loupe API to generate quality assessment reports based on those requirements.

References

1. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality Assessment for linked Data: A Survey. Semantic Web **7**(1) (2016) 63–93
2. Defeo, J.A., Juran, J.M.: Juran's Quality Handbook: The Complete Guide to Performance Excellence. 6 edn. McGraw-Hill Education (6 2010)
3. Olson, J.E.: Data Quality: The Accuracy Dimension. 1 edn. Morgan Kaufmann (1 2003)
4. Rahm, E., Do, H.H.: Data Cleaning: Problems and Current Approaches. IEEE Data Eng. Bull. **23**(4) (2000) 3–13
5. Mihindukulasooriya, N., Rico, M., García-Castro, R., Gómez-Pérez, A.: An analysis of the quality issues of the properties available in the spanish dbpedia. In: Conference of the Spanish Association for AI, Springer (2015) 198–209
6. Mihindukulasooriya, N., Rizzo, G., Troncy, R., Corcho, O., Garcıa-Castro, R.: A Two-Fold Quality Assurance Approach for Dynamic Knowledge Bases: The 3cixty Use Case. In: Proceedings the 1st International Workshop on Completing and Debugging the Semantic Web. (2016) 1–12
7. Böhm, C., Naumann, F., Abedjan, Z., Fenz, D., Grütze, T., Hefenbrock, D., Pohl, M., Sonnabend, D.: Profiling Linked Open Data with ProLOD. In Haas, L., ed.: Proceedings of the 2nd International Workshop on New Trends in Information Integration, IEEE (2010) 175–178
8. Mäkelä, E.: Aether–generating and viewing extended void statistical descriptions of rdf datasets. In: European Semantic Web Conference, Springer (2014) 429–433
9. Khatchadourian, S., Consens, M.P.: ExpLOD: Summary-based exploration of interlinking and RDF usage in the Linked Open Data cloud. In: ESWC. (2010) 272–287
10. Spahiu, B., Porrini, R., Palmonari, M., Rula, A., Maurino, A.: ABSTAT: Ontology-Driven Linked Data Summaries with Pattern Minimalization. In: ESWC (Satellite Events) 2016, Springer (2016) 381–395